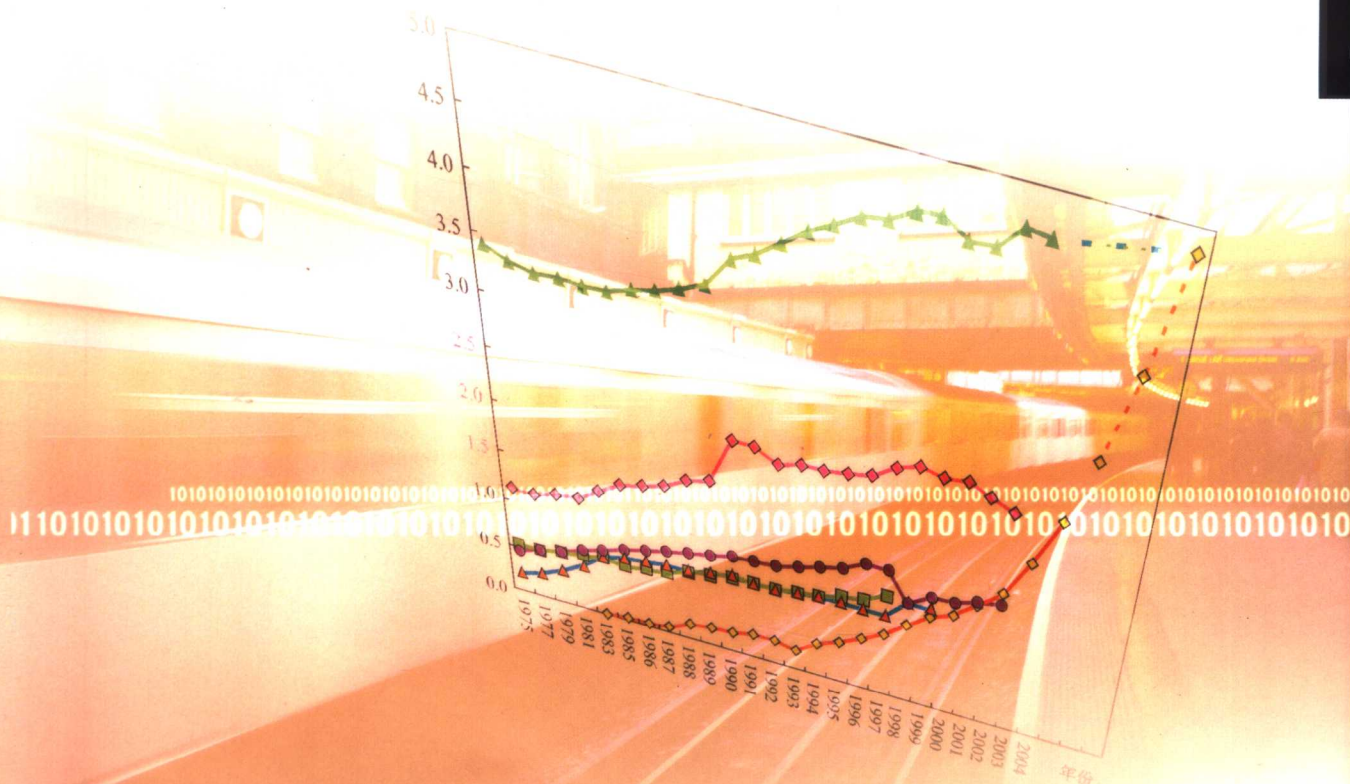


高等学校文科计算机课程系列教材

人文社科信息处理 ——方法与技术

□ 主编 黄都培



高等教育出版社
Higher Education Press

高等学校文科计算机课程系列教材

人文社科信息处理 ——方法与技术

主编 黄都培

高等教育出版社

内容简介

本书编写的目的旨在系统全面地介绍人文社会科学信息处理的方法和技术。本书共分9章,首先给出了两个研究案例,分别为描述性研究和解释性研究,使读者对社会研究有一个清晰形象的认识。然后论述了选择研究课题以及对研究课题进行分析设计的关键技术;在此基础上给出了社会信息量化和操作化的方法以及如何如何进行数据采集、如何对所采集的数据进行预处理;并且结合常用的信息处理工具,如 Excel、SPSS、SAS 等,介绍如何进行信息的分析和处理。最后对信息分析处理技术的发展和应用做了展望。

本书既可作为高校文科专业本科、研究生信息技术应用课程的教材,亦可作为人文社会科学研究人员的辅助参考资料。

图书在版编目(CIP)数据

人文社科信息处理——方法与技术/黄都培主编.

—北京:高等教育出版社,2006.4

ISBN 7-04-018477-X

I. 人... II. 黄... III. 社会科学—信息处理

IV. C37

中国版本图书馆 CIP 数据核字(2006)第 003429 号

策划编辑	孙惠丽	责任编辑	俞丽莎	封面设计	王凌波	责任绘图	朱 静
版式设计	胡志萍	责任校对	王效珍	责任印制	杨 明		

出版发行 高等教育出版社
社 址 北京市西城区德外大街4号
邮政编码 100011
总 机 010-58581000

经 销 蓝色畅想图书发行有限公司
印 刷 北京四季青印刷厂

开 本 787×1092 1/16
印 张 21
字 数 470 000

购书热线 010-58581118
免费咨询 800-810-0598
网 址 <http://www.hep.edu.cn>
<http://www.hep.com.cn>
网上订购 <http://www.landaco.com>
<http://www.landaco.com.cn>
畅想教育 <http://www.widedu.com>

版 次 2006年4月第1版
印 次 2006年4月第1次印刷
定 价 26.00元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换。

版权所有 侵权必究

物料号 18477-00

前 言

如今的现代信息技术和方法不仅是科技人员手中的有力武器,同时其他应用也越来越广泛,深入地运用到人文社会科学的研究中,成为辅助人文社会科学研究必不可少的工具。在社会科学研究中,我们每天都要面对大量的各种各样的数据,要使这些数据转变为对人类做决策有帮助的及时、准确、全面、有用的信息,就必须借助现代化的信息处理技术和方法,对这些数据进行收集、存储、处理和分析。

本书依据系统工程的方法,根据人文社会科学研究的基本流程,对社会学科信息处理和分析各阶段的关键技术和方法做了阐述,并结合大量的案例,从实践的角度向读者全面系统地介绍了开展一项社会研究所用到的计算机软件工具及其在研究各个阶段的作用,通俗、详尽地讲述了如何将抽象的、非结构化的、复杂模糊的客观世界中的社会研究问题转化为机器世界中结构化、可操作、易处理的问题,以便能够使用计算机进行处理和分析。本书旨在向广大读者展现一个系统且实用的人文社会科学信息处理技术方法的全貌。

鉴于社会经济现象之间相互联系、相互制约、纷繁多变,导致人文社会科学中的信息带有极大的主观性、复杂性和不确定性,这就决定了人文社会科学的信息处理具有处理量大而宽泛、信息采集与转换复杂、数据建模和系统设计难度大等特点。我们深感在人文社会科学信息处理这一边缘领域的知识贫乏、才疏学浅,因此,在本书的撰写过程中,我们参考并引用了许多现有的社会统计调查资料和专家学者们的研究成果,力求将自然科学的技术和方法运用于社会学科的研究之中,并在这个领域中有所创新,找出其中信息处理的基本规律。

在此,我们衷心感谢浙江警官学院黄兴瑞教授及其同事们发表的“少年初犯预测研究”以及中国互联网络信息中心(CNNIC)撰写的“第十三次中国互联网络发展状况统计报告”,为本书的编写提供了非常实用的原始案例资料和有利的帮助支持;还要感谢南开大学留美归国社会学博士张光教授的热情指导,使我们在社会学科知识的海洋里受益匪浅,能够在边学习、边实践中完成了这本书的编写工作。

本书由黄都培主持编写,参加策划和编写的还有杨小平、张光、焦敏、刘刚、叶向等。曹卫宁、沈舒对本书提出了十分中肯和建设性的意见,在此表示感谢。由于作者在社会学科领域的认识粗浅,在编写中难免出现许多专业和技术方面的错误,我们诚恳地希望有关专家、学者涵谅,并予以批评指正。

编 者

目 录

1 人文社会科学研究案例	1	3.1.3 数据分类	68
1.1 人文社会科学信息处理综述	1	3.2 社会信息的量化和操作化	69
1.1.1 人文社会科学信息处理的 特征和方法	1	3.2.1 量化与操作化的含义与作用	69
1.1.2 人文社会科学研究的基本流程	3	3.2.2 定量和操作的基本要素	71
1.1.3 本书的编写目的和思路	5	3.3 研究课题的深入	72
1.2 人文社会科学研究案例	6	3.3.1 课题的分解与明确化	72
1.2.1 社会调查研究示例——中国互联 网络发展状况统计报告(节选)	6	3.3.2 提出研究命题与假设	75
1.2.2 社会实证研究示例——少年初 犯预测研究	38	3.3.3 设计测量分析指标与量表	78
1.3 小结	47	3.4 小结	83
思考与习题	47	思考与习题	83
2 开始一项社会科学研究	48	4 获取社会科学研究数据	85
2.1 研究课题的准备	48	4.1 数据采集	85
2.1.1 研究课题的来源	48	4.1.1 收集数据的途径	85
2.1.2 明确研究课题的性质	49	4.1.2 制定调查方案	88
2.1.3 选择研究课题的标准	52	4.1.3 调查的基本方式	89
2.1.4 进行文献回顾和分析	56	4.2 数据抽样	91
2.2 课题维度的设计	57	4.2.1 抽样的概念	91
2.2.1 考察研究的对象	57	4.2.2 抽样的类型	92
2.2.2 设定研究的时间域	58	4.2.3 抽样的流程	96
2.2.3 选择合适的研究方式	59	4.3 调查问卷	99
2.3 课题的申请与论证	61	4.3.1 问卷的基本结构	99
2.3.1 课题的设计论证	61	4.3.2 设计提问项	102
2.3.2 研究的技术和方法	62	4.3.3 构造答题项	103
2.3.3 课题的组织与实施	63	4.3.4 问卷的检验与试用	105
2.4 小结	65	4.4 小结	106
思考与习题	65	思考与习题	107
3 从概念世界到信息世界	66	5 研究分析前的数据准备	108
3.1 社会问题的数据描述	66	5.1 数据质量检验	108
3.1.1 定义概念	66	5.1.1 数据质量与误差	109
3.1.2 变量与常量	67	5.1.2 资料审核与录入	110
		5.1.3 数据检查与清理	111
		5.2 整理数据文件	120
		5.2.1 打开各类格式数据文件	120

5.2.2 数据文件转置	121	7.1.2 个案示例	195
5.2.3 数据排序	122	7.1.3 多重比较	197
5.2.4 数据文件拆分	124	7.1.4 双因素方差分析与多因素方差 分析	198
5.2.5 数据文件合并	125	7.2 回归分析	202
5.3 数据预处理	126	7.2.1 线性回归	204
5.3.1 数据的排秩	126	7.2.2 个案示例	206
5.3.2 自动重编码	127	7.2.3 逐步回归分析	210
5.3.3 派生新变量	128	7.2.4 曲线拟合	213
5.3.4 数据的筛选	130	7.3 聚类分析	216
5.3.5 数据分组归类	133	7.3.1 系统聚类法	217
5.3.6 数据统计计数	136	7.3.2 个案示例	220
5.3.7 数据分类汇总	138	7.3.3 动态聚类法	222
5.3.8 数据加权处理	139	7.4 判别分析	226
5.4 小结	141	7.4.1 距离判别	228
思考与习题	141	7.4.2 Bayes 判别	229
6 社会研究中常用的软件工具	143	7.4.3 Fisher 判别	230
6.1 电子表格软件 Excel 简介	143	7.4.4 个案示例	231
6.1.1 Excel 的特点及基本功能	143	7.4.5 逐步判别	236
6.1.2 Excel 2003 丰富的分析工具	149	7.5 小结	240
6.2 统计分析软件 SPSS 简介	155	思考与习题	240
6.2.1 SPSS 的特点及基本功能	155	8 案例分析与研究报告	246
6.2.2 SPSS 13.0 统计分析功能	161	8.1 银行雇员应用实例	246
6.3 统计分析系统 SAS 简介	175	8.1.1 案例介绍	246
6.3.1 SAS 的特点及基本功能	175	8.1.2 SPSS 统计分析	248
6.3.2 SAS 9.0 数据分析与规划 工具	178	8.2 Excel 在调查问卷统计分析中的应用 案例	269
6.4 数据库管理系统 SQL Server 2000 简介	183	8.2.1 调查问卷统计分析步骤	269
6.4.1 SQL Server 2000 的特点与 功能	183	8.2.2 案例说明	270
6.4.2 SQL Server 2000 的数据库 管理	184	8.2.3 问卷整理及数据录入	271
6.4.3 SQL Server 2000 分析服务	187	8.2.4 基本统计分析	275
6.5 小结	191	8.3 研究报告	294
思考与习题	192	8.3.1 研究报告的基本分类	294
7 社会研究中的常用统计方法	193	8.3.2 研究报告的撰写格式	295
7.1 方差分析	193	8.4 小结	299
7.1.1 单因素方差分析	193	思考与习题	299
7.1.2 个案示例	195	9 信息分析处理技术应用展望	300
7.1.3 多重比较	197	9.1 联机分析处理	300
7.1.4 双因素方差分析与多因素方差 分析	198		
7.2 回归分析	202		
7.2.1 线性回归	204		
7.2.2 个案示例	206		
7.2.3 逐步回归分析	210		
7.2.4 曲线拟合	213		
7.3 聚类分析	216		
7.3.1 系统聚类法	217		
7.3.2 个案示例	220		
7.3.3 动态聚类法	222		
7.4 判别分析	226		
7.4.1 距离判别	228		
7.4.2 Bayes 判别	229		
7.4.3 Fisher 判别	230		
7.4.4 个案示例	231		
7.4.5 逐步判别	236		
7.5 小结	240		
思考与习题	240		
8 案例分析与研究报告	246		
8.1 银行雇员应用实例	246		
8.1.1 案例介绍	246		
8.1.2 SPSS 统计分析	248		
8.2 Excel 在调查问卷统计分析中的应用 案例	269		
8.2.1 调查问卷统计分析步骤	269		
8.2.2 案例说明	270		
8.2.3 问卷整理及数据录入	271		
8.2.4 基本统计分析	275		
8.3 研究报告	294		
8.3.1 研究报告的基本分类	294		
8.3.2 研究报告的撰写格式	295		
8.4 小结	299		
思考与习题	299		
9 信息分析处理技术应用展望	300		
9.1 联机分析处理	300		

9.1.1 OLAP 概述	300	9.2.3 数据挖掘技术	312
9.1.2 OLAP 分类	302	9.2.4 数据挖掘过程	315
9.1.3 OLAP 多维分析	305	9.2.5 数据挖掘工具及展望	316
9.1.4 OLAP 工具及展望	307	9.3 小结	324
9.2 数据挖掘	308	思考与习题	324
9.2.1 数据挖掘概述	308	参考文献	325
9.2.2 数据挖掘功能	310		

1

人文社会科学研究案例

人文社会科学是以人的社会存在为研究对象,以探索、揭示人的本质和人类社会发展规律为目的,它是关于人自身的学说或者理论体系,是对人的存在、本质、价值和发展等问题的研究,同时也是关于人类社会的学说或者理论体系,它以人的共同体的经济活动、政治活动、精神文化活动等社会现象为研究对象。作为一门科学,人文社会科学同样可以采用各科学普遍采用的研究方法。例如,不仅可以采用解释的方法,还可以采用自然科学的实证研究方法,将定性研究与定量研究相结合。

1.1 人文社会科学信息处理综述

人文社会科学所研究的是社会现实中的人、人的行为以及与人有关的社会现象和事物,这些研究对象本身就带着极大的主观性、复杂性和不确定性,使得社会科学研究具有许多不同于自然科学研究的特点和难点。因此人文社会科学信息处理也具有了自己独有的特性。本节将主要讲述人文社会科学信息处理的特征和方法,同时还将阐述进行人文社会科学研究的基本流程。

1.1.1 人文社会科学信息处理的特征和方法

1. 人文社会科学信息处理的特点

人文社会科学领域的信息具有哪些特色?其研究处理方法究竟与自然科学领域所使用的有何不同?这是在开展一项社会研究课题前必须明确的事情。由于人文社会科学本身带有极大的主观性、复杂性和不确定性,使得人文社会科学信息处理也具有较大的复杂性和难度。经过大量实际考察,人文社会科学的信息处理具有如下的特点:

(1) 信息处理量大、范围广

一项社会调查、一个研究课题可能需要收集从历史到现今、从国内到国外、从经济到伦理、从社会到家庭等许许多多、方方面面的信息,这就要求信息处理系统的容量大,处理速度快。

(2) 信息采集与转换工作复杂

因为人文社会科学的信息大部分来自于问卷调查、现场勘察和历史资料中的数据,这些数据大都是以文字、表格、声音、图片等形式表现的多媒体数据,所以在进入信息分析处理之前,首先应该进行数据的收集、存储、转换和预处理等工作。

(3) 信息系统分析设计难度大

人文社会科学的信息基本属于半结构化和非结构化的社会信息,这些信息具有非客观性、模糊性和复杂性,所以必须把大量的精力和时间投入到信息处理的前期——用户需求分析、可行性论证、数据收集整理、问题的分解量化、系统的逻辑分析及原型设计中,而这些工作比自然科学结构化信息的处理要繁琐、艰难得多。

(4) 数据模型复杂

由于人文社会科学中的信息受多种因素影响,社会经济现象之间相互联系、相互制约,因此,一般要求所描述的数据模型必须结构化,避免不相容,这就使得数据模型变得复杂、数据信息之间存在着多种制约关系。

(5) 具有灵敏的环境性

人文社会科学的信息涉及的是上层建筑、文化教育、社会伦理领域,它是整个社会大环境的“晴雨表”,因此,对这门学科的研究和处理会受到国家政治、经济、文化、伦理等社会环境条件及发展变化的干扰和制约。

综上所述,人文社会科学的信息处理与分析是一项长期的、复杂的、动态的、大规模的社会系统工程,对其研究和开发也必须采用既适用又有效的方法,这就是系统工程的方法。

2. 人文社会科学信息处理的方法

所谓系统工程的方法,实际上是一种组织管理技术,是研究大规模、复杂系统的一种科学思想和方法。我国著名科学家钱学森教授指出:“系统工程是组织管理系统的规划、研究、设计、制造、试验和使用的科学方法,是一种对所有系统都具有普遍意义的方法”。

日本学者三浦雄武也曾说:“系统工程的目的是研制一个系统,而系统不仅涉及工程学的领域,还涉及社会、经济和政治等领域,所以为了适当地解决这些领域中存在的问题,除了需要某些纵向技术以外,还要有一种技术从横向把它们组织起来,这种横向技术就是系统工程,也就是研制系统所需的思想、技术、方法和理论等体系化的总称。”

用简单的叙述来概括系统工程方法的要点如下:

- ① 系统的思想——整体考虑研究对象,按步骤分层、分块分析问题。
- ② 定量的方法——数字、程序、模型求解问题。
- ③ 信息技术的应用——现代化科学技术手段处理问题。

系统工程的方法强调用系统化、工程化、规范化的思想来看待和解决问题。系统工程的思想和方法就是要求人们以整体的、动态的、发展的观点分析实际问题,把任何一个待分析处理的事物和问题当做一个系统,综合考虑它的总体目标、各部分间的关系、基本功能、处理逻辑、分析模型、输入/输出以及外部环境等诸多因素的影响,合理地划分系统的层次和联系,找出解决问题的规律及方法,为计算机处理与分析打好基础。

与过去的人工信息系统相比,以计算机为主要信息处理工具的现代信息系统是一种速度快、精度高、容量大的人-机系统。若想发挥计算机信息处理的优势,势必要求将待分析的问题数字化、规范化、标准化和过程化,即用科学、统一的方法、模型、标准来组织、转换、存储数据,对信息

进行加工、传递和输出,特别是对于来自社会学科的半结构化和非结构化问题尤其如此。

这里,仅举全国优秀歌手评选的简单例子,来说明社会科学的分析研究同样要遵循系统工程的方法与程序。其评选程序如表 1-1 所示。

表 1-1 全国优秀歌手评选程序

步骤	内 容
1	提出意向,向全国各地及有关单位发出选拔赛通知
2	收集各地推荐的参赛人员名单、资料,了解参赛选手情况
3	对报送的参赛人员审查、筛选、分组(如,按专业和非专业、年龄段)
4	组委会组织专家评委,编制赛事议程,提出评选方案
5	按评判方案分类命题(如,必选、自选曲目,演唱技巧,文化素质,演员台风等),制订评判标准
6	组织比赛,由专家评委为歌手打分
7	将评委的各类评分输入计算机,与评判标准对比分析,统计处理后输出显示
8	根据综合统计评分结果,选出本届大赛优秀歌手

由表 1-1 可以看到对于全国优秀歌手的评选同样遵循了系统工程的方法和技术,按照系统工程的思想进行报名、筛选、比赛、打分以及最终评选出优秀歌手。

由此可知,人文社会科学的信息系统工程是指用计算机科学、数学模型、管理科学以及社会行为科学的原理、方法、技术来指导、研究、分析和处理相关的信息,它为人文社会科学的分析研究提供了极为有用的、合理科学的研究方法。它在自然科学和人文社会科学之间架设了一座沟通的桥梁,为现代科学技术辅助社会学科研究找到了切入点。

对于人文社会科学课题的研究与开发,按照系统工程的要求,必须把握社会需求,确立研究目标,制订总体规划,开展可行论证,然后根据系统工程给予的研究开发方法,进行设计、分析、研究和开发。

1.1.2 人文社会科学研究的基本流程

作为一种系统的、科学的认识活动,人文社会科学的信息处理和分析研究也遵循着一套比较固定的程序。从通常情况看,可以将分析研究的过程分为以下几个阶段:

- ① 选择研究课(问)题。
- ② 设计研究方案和准备分析工具。
- ③ 数据资料的收集和整理。
- ④ 问题的量化和可操作化。
- ⑤ 问题的处理与统计分析。
- ⑥ 分析结果的解释与报告。

当然,在许多社会实证研究中,还要加入:

- ① 文献回顾和分析。
- ② 提出理论假设。
- ③ 检验并分析假设等。

在实际研究过程中,上述步骤可能交叉重叠,例如,研究者往往是在做文献回顾和分析的过程中,才明确自己的研究课题。但是,这些差异无关宏旨,关键在于能够反映社会科学研究的规律。

1. 选择课题阶段

对于人文社会科学的信息分析来说,选择一个合适的研究课题是整个社会研究工作的第一步,同时也是关键性的一步;研究课题的选择是,在一定程度上决定了整个研究工作的成效。研究课题一旦确定,整个研究活动的目标和方向也就随之确定。有众多的因素决定和制约着研究课题的选择。这些因素包括研究者的研究兴趣、学术水平、文化素养、生活阅历、观察角度,也包括他所处的社会环境、所具备的客观条件等。

选择课题阶段的主要任务包括两个方面:一是选取研究主题,即从现实生活中存在的大量现象、问题和领域中,根据研究者的兴趣、需要与动机确定研究主题;二是构建研究课题的模型,即进一步明确对这一课题研究的范围、性质、结构,集中研究的焦点,将最初比较含糊、比较笼统、比较宽泛的研究领域和研究现象具体化、精确化、层次化,将其转化为既有价值又有新意,同时还切实可行的研究课题。这一过程既是将课题从一般到特殊的筛选过程,也是从模糊到清晰的认知过程。

2. 研究设计阶段

就像建设大厦之前必须进行建筑设计一样,要保证一项社会研究工作的顺利进行以及研究目标的圆满实现,也必须进行周密的研究设计。如果把选择的研究课题比喻为规划大厦的总体蓝图,那么研究设计阶段的全部工作就可以理解成为实现蓝图所开展的具体规划和准备工作。社会研究的设计包含了从研究目的、研究用途、研究方式、分析对象,直到具体研究方案等的策略、方法、技术及各个方面,还包括了对研究所依赖的测量工具、测试方法和信息收集、处理和分析工具,如问卷、量表、分析软件等的准备。在设计方案基本确定后,应提交一份包括预期目标、现状分析、研究方法、采用技术、成果形式、时间进度、材料经费等内容的可行性论证报告。

3. 数据收集阶段

这个阶段的任务是贯彻研究设计中确定的策略,按照设计中确定的方式、方法和技术进行数据的收集和整理工作。其主要特点是:研究者往往要深入实地接触被研究对象,开展调查研究或设计出实验环境,通过不同的调查和测量方法,收集大量的第一手和第二手资料。因为数据是信息分析处理的原料,没有来源于客观实际的数据就不会产生任何具有实用价值的研究成果。第一手(原始)数据收集是指调查的当时当地,从客观实际中直接取出数据,并用特定的技术手段在介质上记录下来;第二手(二次)数据收集则是指收集已记录在某种介质上、与所描述的实体在时间与空间上已分离开的数据,如文献资料。这一步工作的关键是收集的数据的准确性和真实性。

4. 数据筛选阶段

数据筛选是社会学科信息处理过程中不可忽视的阶段。因为自客观世界收集的数据往往是大量的、混杂的、不规范的、非结构化的,这将给下一步的分析和处理带来极大的困难,必须事先对它们进行筛选、整理,转换为信息系统所能描述、存储、处理的数据。这个阶段涉及数据的抽样方法、数据格式、存储方式、编码转换、检查校验等一系列设计的工作,既有对原始数据资料的清理、转化和录入计算机等工作;也有对原始文字资料、图片资料、音像资料等的整理、分类和加工,以保证后期数据分析处理的顺利进行。在这一阶段中,投入的人力最多,工作量极大,因此,需要很好地组织和管理。另外需要注意的是,由于社会现象的复杂性,或者由于现实条件的变化,事前所考虑的研究设计往往会在某些方面与现实之间存在一定的距离和偏差,这就需要根据实际情况进行修整或弥补,发挥研究者的灵活性和主动性。

5. 数据分析阶段

数据分析阶段也称为研究阶段。这一阶段的主要任务是对前一阶段收集筛选的资料进行系统的归类、统计、计算和分析。从现实生活中所得到的众多信息和资料,要经过研究者的加工和处理,才能最终变成研究的结果和结论。这里既包括对数据资料进行的各种数值计算、统计等定量分析,也包括对定性资料进行的综合、归纳和解释分析。目前,利用 Excel、SAS、SPSS 等专门的计算机分析处理工具以及数据挖掘工具辅助研究分析,可以大大简化分析过程,提高处理效率,为人们提供精确、优质的信息。

6. 得出结果阶段

此阶段的工作是:撰写研究报告,评估研究质量,交流研究成果。研究报告是一种以文字和图表等形式来系统地、集中地、规范地反映整个研究工作所得到的结果,它是社会研究成果的集中体现。而撰写研究报告也可以说是对整个社会研究工作进行全面的总结。从研究的目的、方式,到资料的收集、分析方法以及研究得出的结论、研究成果的质量,都要在研究报告中进行总结和反映。同时,还要将社会研究的成果以不同的形式应用到社会实践中,真正发挥社会研究在认识社会现象、探索社会规律中的巨大作用。

1.1.3 本书的编写目的和思路

社会经济现象之间的相互联系、相互制约使得人文社会科学的信息带有极大的主观性、复杂性和不确定性。这就决定了人文社会科学的信息处理具有量大而宽泛、信息采集与转换复杂、信息系统设计难度大、数据模型复杂等特点。本书依据系统工程的方法,按照人文社会科学研究的基本流程,对人文社会科学信息处理和分析各阶段的关键技术和方法做了阐述;并结合大量的案例,从实践角度论述了如何进行高效、合理的信息处理和分析。

本书编写的目的旨在系统全面地介绍人文社会科学信息处理的方法和技术。本书首先给出两个研究案例,分别为描述性研究和解释性研究,使读者对社会研究有一个清晰、形象的认识。然后论述选择研究课题以及对研究课题进行分析设计的关键技术;在此基础上给出了社会信息量化和操作化的方法以及如何进行数据采集、如何对所采集的数据进行预处理。并且结合常用

的信息处理工具,如 Excel、SPSS、SAS 等给出进行信息的分析和处理的方法。最后对信息分析处理技术的发展和应用进行展望。

1.2 人文社会科学研究案例

人文社会科学的研究可以归纳为以下三类:探索性研究、描述性研究和解释性研究。描述性研究通常是指社会调查研究,用于对总体情况或某种现象的描述。解释性研究也称实证研究,用于解释和说明问题及现象发生的原因,发生变化的内在规律以及之间的关系。本章分别从社会调查研究和实证研究给出两个实例:“中国互联网络发展状况统计报告”和“少年初犯预测研究”。“中国互联网络发展状况统计报告”分别论述了调查背景、调查内容、相关说明、调查结果、调查方法,最后给出了中国互联网络发展状况分析报告,使读者可以完全掌握和领悟调查报告的方法、技巧和精华。“少年初犯预测研究”是一个典型的实证研究,分别从社会因素和心理因素做了少年初犯预测研究。

1.2.1 社会调查研究示例——中国互联网络发展状况统计报告(节选)

第十三次中国互联网络发展状况统计报告^①

CNNIC

中国互联网络信息中心

目 录

- 第一部分 调查背景
- 第二部分 调查内容
- 第三部分 相关说明
- 第四部分 调查结果
 - 一、中国互联网络发展的宏观概况
 - 二、互联网用户行为意识调查结果
 - (一) 用户个人信息(删节)
 - (二) 用户对互联网的使用情况及满意度
 - (三) 用户对互联网热点问题的回答(删节)
- 第五部分 调查方法
 - 一、域名数、网站数调查
 - 二、网上联机调查

^① 本案例摘自于中国互联网络信息中心 CNNIC 网站: <http://www.cnnic.net.cn>

三、网下抽样调查

第六部分 调查支持单位(删节)

第七部分 中国互联网络发展状况分析报告(部分删节)

附录一 中国香港地区互联网络发展状况统计报告(删节)

附录二 中国澳门地区互联网络发展状况统计报告(删节)

第一部分 调查背景

我国互联网络上网计算机数、用户人数、用户分布、信息流量分布、域名注册等方面情况的统计信息,对国家和企业动态掌握互联网络在我国的发展情况,提供决策依据有着十分重要的意义。

1997年,经国家主管部门研究,决定由中国互联网络信息中心(CNNIC)联合互联网络单位来实施这项统计工作。为了使这项工作制度化、正规化,从1998年起CNNIC决定于每年1月和7月发布“中国互联网络发展状况统计报告”。统计报告发表后,受到各个方面的重视,被国内外用户广泛引用。本次为CNNIC第十三次调查。

值得说明的是,该项工作得到了信息产业部等国家主管部门的大力支持,各互联网单位、调查支持网站以及媒体等也对CNNIC给予了支持与配合,使得中国互联网络发展状况统计调查工作得以顺利进行。

第二部分 调查内容

第十三次中国互联网络发展状况统计调查的主要内容为:

(一) 中国互联网络发展的宏观概况

包括:我国上网计算机数

我国上网用户人数

CN下注册的域名数及其地域分布情况

WWW站点数及其域名、地域分布

我国国际出口带宽总量

我国IP地址总量

(二) 中国互联网络用户的相关情况

包括:用户基本特征(如性别、年龄、婚姻状况、文化程度等)

用户对互联网的使用情况及满意度

用户对互联网热点问题的回答

第三部分 相关说明

第十三次互联网发展状况统计报告相关说明:

1. 网民(互联网用户):CNNIC对网民的定义为:平均每周使用互联网至少1小时的中国

公民。

2. 网站:指有独立域名的 web 站点,其中包括 CN 和通用顶级域名(gTLD)下的 web 站点。此处的独立域名指的是每个域名最多只对应一个网站“WWW. + 域名”。如,对域名 cnnic. net. cn 来说,它只有一个网站 www. cnnic. net. cn,并非它有 whois. cnnic. net. cn、dns2. cnnic. net. cn 等多个网站,它们只被视为网站 www. cnnic. net. cn 的不同频道。

3. 上网计算机:指至少有一人通过该台计算机连入互联网络。

4. 除非明确指出,本报告中的数据均不包括中国香港、澳门、台湾地区在内。

5. 本次调查统计数据截止日期为 2003 年 12 月 31 日。

第四部分 调查结果

一、中国互联网络发展的宏观概况

(一) 我国上网用户人数

1. 上网用户总人数为 7 950 万,95% 置信度下的置信区间为【7 704 万,8 196 万】。

2. 按上网方式划分,参见表 1-2。

表 1-2 按上网方式划分(人数)

专线上网用户数	拨号上网用户数	ISDN 上网用户数	宽带上网用户数
2 660 万	4 916 万	552 万	1 740 万

注 1:通过多种方式上网的用户被重复计入各种上网方式中,故各种方式上网用户数之和大于上网用户总数;

注 2:专线上网用户指通过以太网方式接入局域网,然后再通过专线的方式接入互联网的用户;

注 3:宽带上网用户指使用 xDSL、CABLEMODEM 等方式上网的用户。

3. 按地域划分,参见表 1-3。

表 1-3 按地域划分(人数)

	北京	上海	天津	重庆	河北	山西	内蒙
网民数(万)	398.0	431.6	144.6	176.6	289.1	148.8	74.9
占全国网民比例	5.0%	5.4%	1.8%	2.2%	3.6%	1.9%	1.0%
占本省人口比例	28.0%	26.6%	14.4%	5.7%	4.3%	4.5%	3.1%
	辽宁	吉林	黑龙江	江苏	浙江	安徽	福建
网民数(万)	291.5	146.5	226.0	610.9	451.2	183.5	318.2
占全国网民比例	3.7%	1.8%	2.8%	7.7%	5.7%	2.3%	4.0%
占本省人口比例	6.9%	5.4%	5.9%	8.3%	9.7%	2.9%	9.2%

续表

	江西	山东	河南	湖北	湖南	广东	广西
网民数(万)	169.4	626.6	225.7	380.9	265.4	950.2	228.6
占全国网民比例	2.1%	7.9%	2.8%	4.8%	3.3%	12.0%	2.9%
占本省人口比例	4.0%	6.9%	2.3%	6.4%	4.0%	12.1%	4.7%
	海南	四川	贵州	云南	西藏	陕西	甘肃
网民数(万)	39.7	424.3	83.1	166.4	8.6	196.7	122.4
占全国网民比例	0.5%	5.3%	1.1%	2.1%	0.1%	2.5%	1.5%
占本省人口比例	4.9%	4.9%	2.2%	3.8%	3.2%	5.4%	4.7%
	青海	宁夏	新疆				
网民数(万)	19.5	33.3	117.8				
占全国网民比例	0.3%	0.4%	1.5%				
占本省人口比例	3.7%	5.8%	6.2%				

注:计算中所用的各省总人口数为《中国统计摘要》公布的2002年底数据。

4. 除计算机外同时使用其他设备(移动终端、信息家电)上网的用户人数为214万。

(二) 我国上网计算机数

1. 上网计算机总数为3 089万。

2. 按上网方式划分,参见表1-4。

表1-4 按上网方式划分(台数)

专线上网计算机数	拨号上网计算机数	其他方式上网计算机数
595 万	1 945 万	549 万

(三) CN下注册的域名数

1. 域名总数为340 040个。

2. 按域名类别划分,参见表1-5。

表1-5 按域名类别划分(个数)

	AC	COM	EDU	GOV	NET	ORG	行政区域名	二级域名(.CN)
数量	666	140 779	1 915	11 764	16 189	7 369	3 286	158 072
百分比	0.2%	41.4%	0.5%	3.5%	4.7%	2.2%	1.0%	46.5%

各类域名所占的比例如下图所示,参见图1-1。

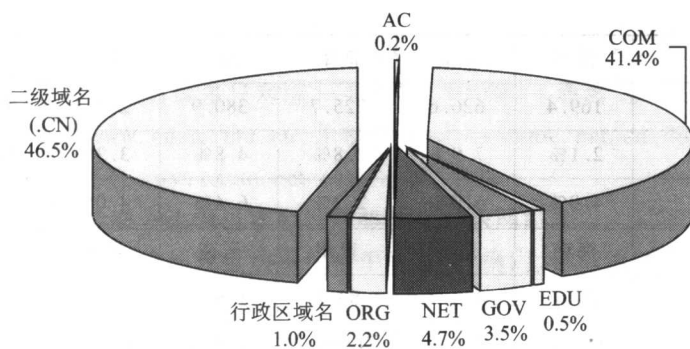


图 1-1 各类域名所占比例图

3. 按地域划分(不含 EDU), 参见表 1-6。

表 1-6 按地域划分(不含 EDU)的域名(个数)

	北京	上海	天津	重庆	河北	山西	内蒙
域名数量	84 144	28 940	5 366	3 651	6 135	1 684	1 942
百分比	24.9%	8.5%	1.6%	1.1%	1.8%	0.5%	0.6%
	辽宁	吉林	黑龙江	江苏	浙江	安徽	福建
域名数量	9 793	3 659	3 332	20 587	19 780	3 649	11 395
百分比	2.9%	1.1%	1.0%	6.1%	5.8%	1.1%	3.4%
	江西	山东	河南	湖北	湖南	广东	广西
域名数量	2 311	13 429	4 898	6 131	3 216	47 322	2 751
百分比	0.7%	4.0%	1.4%	1.8%	0.9%	14.0%	0.8%
	海南	四川	贵州	云南	西藏	陕西	甘肃
域名数量	1 150	7 380	1 269	3 539	624	4 030	1 274
百分比	0.3%	2.2%	0.4%	1.0%	0.2%	1.2%	0.4%
	青海	宁夏	新疆	海外			
域名数量	348	1 022	2 579	30 795			
百分比	0.1%	0.3%	0.8%	9.1%			

(四) WWW 站点数(包括.CN、.COM、.NET、.ORG 下的网站):

- WWW 站点总数为 595 550 个。
- 按所属域名类别划分, 参见表 1-7。